



**BRC4 Meeting
December 7, 2006
University of Alabama - Birmingham**

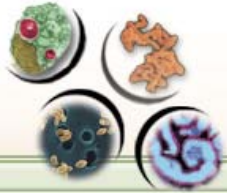
BioHealthBase Yesterday, Today, and Tomorrow

**The BioHealthBase Bioinformatics Resource Center for Biodefense and
Emerging/Re-emerging Infectious Disease**

**Kevin J. Biersack
Northrop Grumman Health Solutions
Rockville, MD**



BioHealthBase: Home Page



BIOHEALTHBASE HOME

ORGANISMS

FAQ

FEEDBACK

Publications

BRC News and Events

Mission

Release Notes

Data Loads

Related Links

Science Support

Organisms



Influenza Virus



Francisella tularensis



Mycobacterium



Microsporidia



Giardia

Genomes in BioHealthBase

Species-based

Organism	Kingdom	# Strain
<i>Mycobacterium leprae</i>	Bacteria	1
<i>Mycobacterium avium</i>	Bacteria	1
<i>Francisella tularensis</i>	Bacteria	3
<i>Mycobacterium</i>	Bacteria	1
<i>Mycobacterium tuberculosis</i>	Bacteria	2
<i>Mycobacterium bovis</i>	Bacteria	1
<i>Encephalitozoon cuniculi</i>	Fungi	1
Influenza C Virus	Virus	154
Influenza A Virus	Virus	7221
Influenza B Virus	Virus	1136

Kingdom-based

Kingdom	# Species	# Strain
Virus	3	8511
Fungi	1	1
Bacteria	6	9
Total	10	8521

What's New

September 29, 2006 - Release 2.2

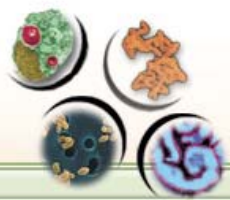
New/Updated Data

- Added NetCTL epitope predictions for new influenza sequences
- Updated influenza IEDB curated epitopes
- Updated influenza concatenated sequence A/Puerto Rico/8/34
- Added SNP data for influenza concatenated reference sequence A/Puerto Rico/8/34
- Added PSIPRED protein secondary structure predictions for *Francisella Schu 4* proteins
- Added protein subcellular localization predictions for *Francisella tularensis* FSC 198 and *Mycobacterium* sp. MCS proteins
- Added Glimmer predictions for *Francisella tularensis* FSC 198 and *Mycobacterium* sp. MCS genomes
- Added BlastP:Swiss-Prot alignments for *Francisella tularensis* FSC 198 and *Mycobacterium* sp. MCS proteins
- Updated UniProt data

Our Mission

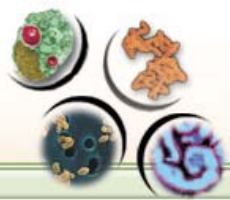
The primary mission of the BioHealthBase system is to assist scientific researchers in their development of vaccines, therapeutics, and diagnostics. The National Institute of Allergy and Infectious Disease (NIAID) Division of Microbiology and Infectious Diseases (DMID) recognizes the challenge posed by bioterrorism, the emergence of disease due to drug-resistant variants of etiologic organisms. DMID has envisioned a consortium of Bioinformatics Resource Centers (BRCs) for Biodefense and Emerging/Re-emerging Infectious Diseases that will provide information technology (IT) support for experimental studies of pathogenic organisms that could be used for biowarfare and bioterrorist activities, many of which also pose an ongoing threat to public health. The BRCs will provide both central repositories for a wide variety of scientific data on these pathogenic microorganisms and a platform for software tools that support investigator-driven data analysis. A description of the NIAID BRC program can be found at www.niaid.nih.gov/dmid/genomes/brc/default.htm

www.biohealthbase.org



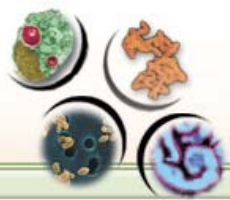
V1.0 Baseline Release – S/W

- Database Search and Display
 - Virus (*Influenza*)
 - Gene, Epitope, Public Database Identifier, Polymorphism
 - Bacteria (*MTB and Tularemia*)
 - Gene Search, Public Database Identifier, Generic locus, Generic GO
 - Retrieve the following detailed gene and protein information for a selected gene:
 - Gene Identification, Gene and feature location, Operon identification (Bacteria Only), Protein identification, Protein domain and motif mappings, Preliminary Protein Localization (Bacteria), Preliminary epitope hits for MHC superfamilies (Influenza), GO Classification, Database cross-reference, Literature References, Data sources (with footnotes)
 - Genomic browsing via the GMOD GBrowse tool (bacteria only)
 - Data download:
 - Genome, gene, and feature sequences in FASTA format
 - Annotation in GFF3 format
- Data Processing (Pre-computes)
 - Semi-automated data processing pipeline (initially, BLAST using the NCBI non-redundant amino acid (NR) database)
 - Creation of consensus sequences for each influenza serological type
 - Influenza frequency of minor allele computation with comparison to the consensus influenza sequences for each serological type
 - Initial protein localization computations for bacterial organisms
 - Preliminary MHC class I epitope prediction for influenza



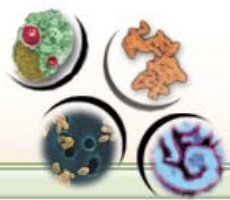
V1.0 Baseline Release - Data

- *Francisella tularensis* genome sequence and annotation
 - *Schu S4* - Provisional RefSeq
- *Mycobacterium tuberculosis* genome sequence and annotation
 - *Mycobacterium tuberculosis* - CDC 1551 - Provisional RefSeq
 - *Mycobacterium tuberculosis* - H37Rv - Provisional RefSeq
 - *M. bovis*, *M. leprae*, and *M. avium* - Provisional RefSeqs
- *Influenza virus* (types A, B, and C) - from NCBI
 - All influenza nucleotide and protein sequences provided to NCBI
- Protein information, integrated with corresponding NCBI annotation and sequences:
 - UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, and Pfam (domains and motifs)
- Generated Data:
 - Protein localization (PSortB, SubLoc, SignalP, DAS - bacteria only)
 - MHC epitopes (NetCTL - *influenza* only)
 - Sequence polymorphism analysis etc. (MUSCLE plus custom software - *influenza* only)
- Pathway Data:
 - BioCyc metabolic pathways (bacteria only)
 - Operons (BioCyc Pathway Tools - bacteria only)
 - Reactome interferon pathway (*influenza* only)



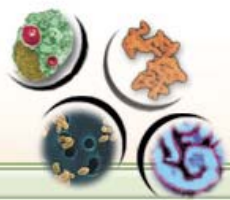
Through V2.2 Release – S/W

- Database Search and Display
 - *Encephalitozoon cuniculi* (Microsporidia)
 - Gene, Locus, Public Database Identifier, GO, and Pfam Domain and Motif
 - Bacteria: Pfam Domain and Motif
 - Immune Epitope Database (IEDB) Epitope Search (*influenza*)
 - Gene Ontology (GO) Keyword Search (bacteria, Microsporidia)
 - User-interactive blast (blastn, blastp, blastx)
 - Blastp alignments (top 10 + detailed report)
 - SNP details, Amino acid consensuses, IEDB validated epitope data display
 - Additional GBrowse tracks
 - Added IEDB Eptiope track, SNPs, Predicted MHC Class I epitopes (*influenza*)
 - Added SNP track and display for the concatenated reference sequence *A/Puerto Rico/8/34*
 - Added PSIPRED Prediction track (*Francisella Schu 4*)
 - Blastp alignments (top 50), Pfam domains/motifs, Glimmer ORFs, 6-frame translation
- Tools: Added user-interactive multiple sequence alignment (MUSCLE)
- New Features
 - Gene Cart allowing users to store and download sequences and annotations
 - Add linkouts from NCBI Entrez gene ID, taxon ID, genome accession, UniProt accession, and other locations.
 - For flu, dynamically listed the appropriate hosts, subtypes, and genes
 - Added additional influenza consensus sequences for non-human hosts. Dynamically populated the subtype list based on host, and the segment list based on subtype
 - Column sorting for search result tables



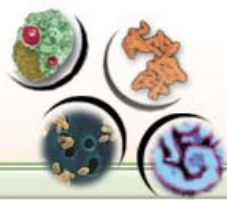
Through V2.2 Release - Data

- New
 - *Influenza* sequences
 - Amino acid consensuses (*influenza*)
 - SNP data and concatenated sequences (*influenza*) *Encephalitozoon cuniculi*
 - *Francisella LVS*, *Schu 4* (protein secondary structure predictions), *FSC 198* (glimmer predictions/protein subcellular localization)
 - *Mycobacterium sp. MCS* (and glimmer predicions/protein subcellular localization)
 - Immune Epitope Database (IEDB) validated epitopes and NetCTL epitope predictions
- Protein information, integrated with corresponding NCBI annotation and sequences, with updates from:
 - UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, Pfam (domains and motifs)
- Blastp alignments for all BioHealthBase proteins mapped against the SwissProt protein database
- Updates – **Monthly**
 - Internal : October 2006
 - **BRC Central (gff3)**: November 2006



Key BioHealthBase Features

- BioHealthBase is an **integrated** resource
 - Interrelates data from NCBI, UniProt, Pfam, and other sources
 - Direct summary and visualization of integrated data
 - Linkouts to source data sites for additional data details
 - Allows “one-stop shopping” for science researchers
- BioHealthBase provides value-added, **pre-computed** data
 - Valuable processed data not available elsewhere
 - Protein structural and functional predictions
 - Sequence alignment and polymorphism analysis
 - Predicted immunological epitopes
- BioHealthBase provides a forum for displaying **collaborative** data
 - Pathways with Reactome; epitopes with Immune Epitope Database
- BioHealthBase emphasizes **host-pathogen interactions**
 - Pathogen effects on host cellular pathways
 - Immune response to pathogen infection
 - Support of research related to vaccine, therapeutic and diagnostic development



BHB Primary Data Sources



Data Load - Microsoft Internet Explorer provided by Northrop Grumman Corporation

File Edit View Favorites Tools Help

Back Forward Search Favorites

Address <http://www.biohealthbase.org/GSearch/statsAutomation.do?decorator=BioHealthBase&pageType=DataLoads> Go Links

BioDefense Public Health Database

BIOHEALTHBASE HOME ORGANISMS FAQ FEEDBACK

Publications
News and Events
Collaborations
Mission
Release Notes
Data Loads
Related Links
Science Support

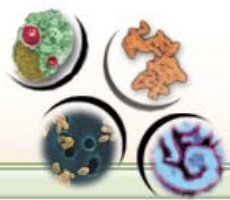
External Data Loads

Organism	Data Type	Data Source	Release/Version #	Release/Download Date	Last Updated Date
Influenza	Sequence/Annotation	NCBI-GenBank	156	08/2006	11/2006
Influenza	Sequence/Annotation	NCBI-RefSeq	19	09/2006	10/2006
Influenza	Curated Epitopes	Immune Epitope Database (IEDB)	N/A	09/2006	09/2006
Bacteria/Microsporidia	Sequence/Annotation	NCBI-RefSeq	20	09/2006	11/2006
MTB	Transposon Insertion Mutants	TARGET	N/A	07/2006	11/2006
All	Swiss-Prot	UniProtKB	50.6	09/2006	09/2006
All	TrEMBL	UniProtKB	33.6	09/2006	09/2006
All	Domains/Motifs	Pfam	20	05/2006	07/2006

Internal Data Loads

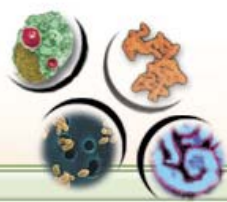
Organism	Data Type	Program	Last Updated Date
Influenza	CTL Epitopes	NetCTL	11/2006
Influenza	SNP Data	MUSCLE, Internal Pipeline	06/2006
Influenza	Consensus Sequence	MUSCLE, Internal Pipeline	06/2006
Influenza	Sequence Similarities	NCBI BlastP	11/2006
Influenza	Concatenated Reference Sequence	Internal Pipeline	10/2006
Bacteria	Protein Subcellular Localization Predictions	PSortB, SignalP, SubLoc, DAS	11/2006
Bacteria	Operon Predictions	BioCyc Pathway Tools	11/2006
Bacteria	Sequence Similarities	NCBI BlastP	11/2006
Bacteria	Metabolic Pathways	BioCyc Pathway Tools	01/2006
Bacteria	ORF Predictions	Glimmer	09/2006
Bacteria	Protein Secondary Structure Predictions	PSIPRED	11/2006
Microsporidia	Sequence Similarities	NCBI BlastP	10/2006

Internet



Existing Collaborations

- **Reactome** - an open source, curated resource of core pathways and reactions in human biology authored by biological researchers with expertise in their fields (**Influenza Infection and Influenza Life Cycle**)
- **TARGET** (Tuberculosis Animal Research and Gene Evaluation Taskforce) BHB Mycobacterium database provides the visualization of insertion coordinates for each of these mutants using GBrowse as well as a direct link to the Target's and CSU's web-page for ordering the mutant strains.
- **BioCyc** - integrating metabolic pathway information using the Mycobacterium tuberculosis (Mtblvcyc) and Francisella tularensis (Frantcyc) pathway databases and using the Pathway Tools and related databases (<http://biocyc.org>) to obtain and update the metabolic pathway information for BHB pathogens.
- **IEDB** — For influenza virus, we are integrating their curated database of immune epitope data and analysis resources for antibody and T cell epitope data and the MHC binding data from a variety of different antigenic sources and immune epitope data. ⁹



Future Collaboration – Influenza Sequence Database (ISD)



ISD: Influenza Sequence Database at LANL - Microsoft Internet Explorer provided by Northrop Grumman Corporation

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail New Window

Address <http://www.flu.lanl.gov/> Go Links >>

The Influenza Sequence Database

Home | Search | Vaccine Selection | Reviews | Links | Contact Us | Help | Report a Bug | Login

About the ISD

The Influenza Sequence Database (ISD) contains all published influenza viral sequences, which have been curated by domain experts to ensure high standards of accuracy and completeness. ISD tools for management of influenza viral sequences and study of the molecular evolution of the influenza virus are designed to expedite the typical tasks of molecular epidemiology. Subscriptions to the ISD are required for full access to tools; information on the purchase of subscriptions can be obtained by emailing flu@lanl.gov.

Login to the database

Welcome to the Influenza Sequence Database.

You have several options for using the ISD.

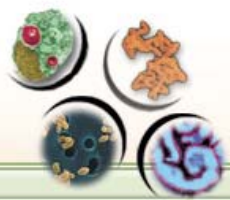
Cost	Services
Free	<ul style="list-style-type: none">• Access to all curated, public data• Download of moderate amounts of data• Access to Vaccine Selections and other Reviews
Subscription	<ul style="list-style-type: none">• All of the above, plus:• All tools, such as BLAST, alignments, inference of phylogenies
Work-for-Others	<ul style="list-style-type: none">• All of the above, plus:• Ability to store private data• Ability to save working sets• Ability to share data with other WFO-users

In addition, we endeavor to obtain contributions from other institutions toward providing free subscriptions to those unable to purchase a subscription. The amount of needs-based support is limited and fluctuates; requests for free subscriptions are fulfilled only if a clear need is evident and funds are available.

Notice to Users: Please cite the Influenza Sequence Database (ISD) in your publications as follows: Macken, C., Lu, H., Goodman, J., & Boykin, L., "The value of a database in surveillance and vaccine selection." in *Options for the Control of Influenza IV*. A.D.M.E. Osterhaus, N. Cox & A.W. Hampson (Eds.) Amsterdam: Elsevier Science, 2001, 103-106.

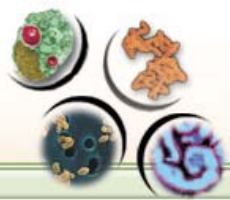
Los Alamos National Laboratory, Operated by [LANL LLC](#) for the US Department of Energy | [Copyright © 2006 LANL LLC](#) | [Disclaimer/Privacy Policy](#)

Internet



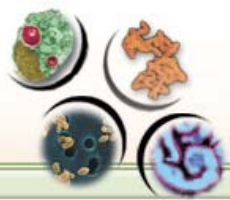
ISD 2-Year Plan

- Interagency Agreement between DOE/LANL and NIH/NIAID
- 2 Year Timeline: January 2007 – December 2008
- Three Phases
- End Result – Integrate LANL ISD content and functions into BHB BRC
- Engage Catherine Macken, one of our SWG members, as a co-PI for the BHB BRC



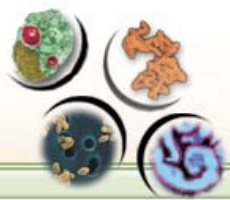
Future directions

- Current plans
 - Continue to enhance pathway data
 - Phylogenetic tree construction
 - **3D protein** chemical structure **visualization**
 - Additional manual curation
 - Add *Giardia lamblia* and *ricinis communis* into BHB site/resource
- Advice and feedback
 - feedback@biohealthbase.org



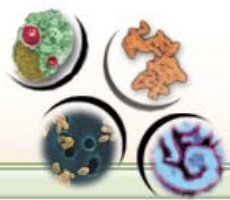
3D Protein Visualization

- Summary Requirements
 - Implement 3D visualization tool
 - Map sequence features onto 3D visualization views
 - Map multiple sequence features onto 3D visualization views
 - Map non-linear sequence features onto 3D visualization views
 - Map 3D protein view as amino acids
 - Enable search in 3D
 - View sequences feature values
 - Visualize 3D structure of composite sequences from multiple sequences
 - Visualize binding
 - Save 3D visualization as movie or photo (if not built in)
 - Create visual shortcuts for popular views
 - Align primary, secondary and 3D structure in a collapsible single view
 - Integrate protein-to-protein binding database data to be able to bind companion proteins to original selected protein in 3D
 - Enable oligomerization or multiple entity binding of monomers



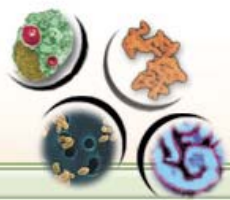
3D Visualization - Functional Requirements

- Display 3D structure of proteins
- Mapping / Highlighting
 - Highlight genome features on 3D structure
 - Highlight multiple genome features
 - Highlight based on annotation characteristic threshold
- Search
 - Search for region based on sequence, annotation
- Output
 - Save publication quality image, movies
- Binding
 - Protein-to-protein
 - Protein-to-small molecules
- Comparison
 - Multiple sequence alignment in 3D
 - Sequence similarity
 - Highlight differences, similarities



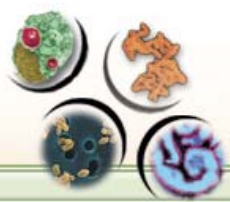
3D Visualization Application Evaluation

- Criteria
 - Web application
 - Open source
 - Cross-platform
 - Browser independent
 - Require no installation
 - Native PDB support
 - Easily integratable
 - Allow highlighting or feature mapping
- Evaluations
 - Jmol
 - Protein Explorer (RasMol)
 - Chime
 - FPV
 - AutoDock
 - Chimera
 - Cn3D
 - MICE
 - QuickPDB
 - VMD
 - Strap

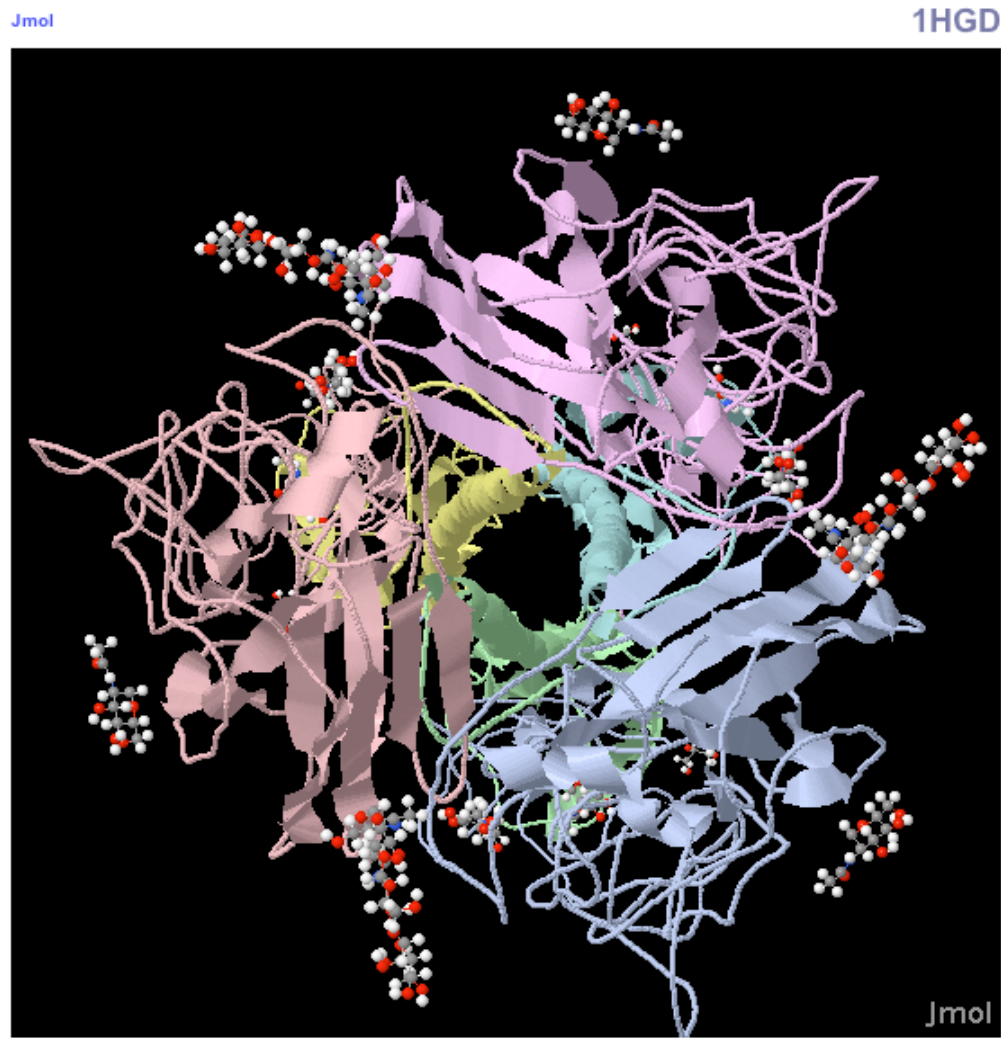


Structural Biology in BHB

- 600 structures from four organisms
- Primarily Influenza and Mycobacterium
- 0 structures for Francisella or E cuniculi
- Viewed in Jmol applet
- Processing offloaded to client side



Influenza *HA trimer binding with sialic acid*





Jmol

1HGD



Jmol

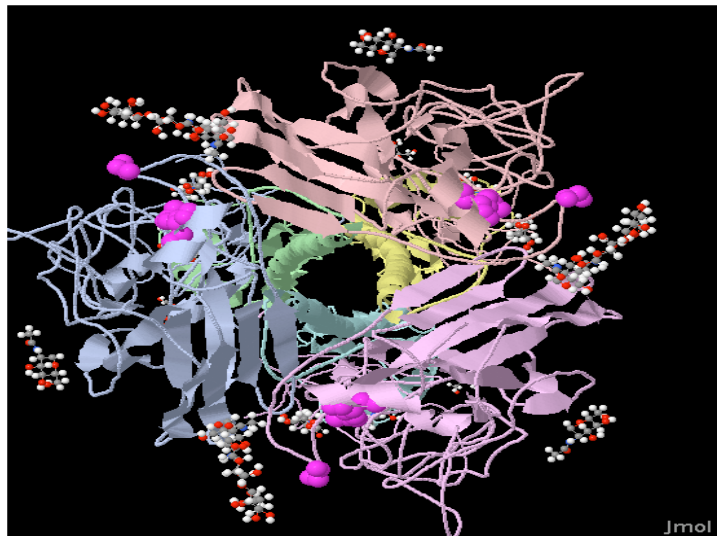
1HGD

hBase
th Database



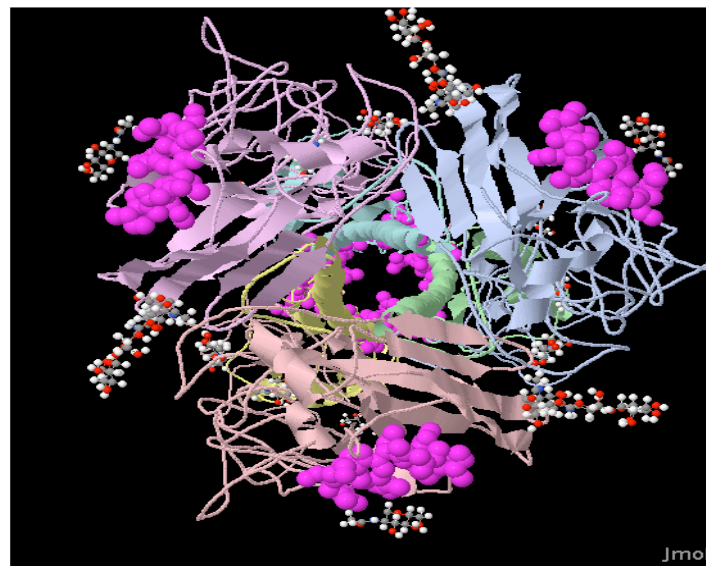
Jmol

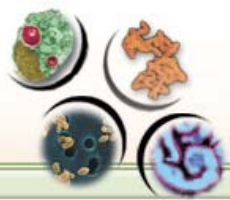
1HGD



Jmol

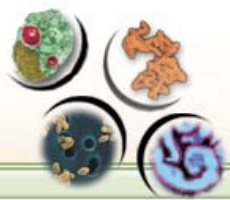
1HGD





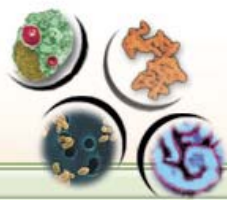
Progress in BHB visualization

- **Present:** (Stock)
 - Load all 600 Structures into GUS 3.5
 - Enable viewing of those structures
- **Near Term:** (Mods)
 - Link to Gbrowse features
 - Search for structures
 - Annotate features
- **Future:** (Custom Algorithms)
 - Substructure searching
 - Threading onto consensus sequence



Acknowledgments

- U.T. Southwestern
 - Burke Squires
 - Feng Luo
 - Shubhada Godbole
 - Jennifer Cai
 - Jyoti Shah
 - Jamie Lee
 - Cathy Spranger
 - Victoria Hunt
- SRI
 - Peter Karp
 - John Pick
- Reactome
 - Marc Gillespie
 - Peter D-Eustachio
- Northrop Grumman
 - Kevin Biersack
 - Ed Klem
 - Carey Gire
 - Jason Lucas
 - Sharmila Vattikuti
 - Sanjeev Kumar
 - Paul Shrabstein
 - Surabhi Sharma
 - Tammie Ajayi
 - Aihui Wang
 - Zuoming Deng
 - Jianjun Wang
 - DeWayne Ejikeme
 - Soumya Sengupta
 - Doug Marcey
- Vecna
 - James Wolowicz
 - Chris Larsen
 - Al Ramsey
- Amar
 - X. Wei



Acknowledgments - Continued

SWG List - Microsoft Internet Explorer provided by Northrop Grumman Corporation

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail W Y Go Links

Address <http://www.biohealthbase.org/GSearch/Swg.jsp?decorator=BioHealthBase>

Release Notes
Data Loads
Related Links
Science Support

University of Texas Southwestern Medical Center

Co-Investigators

Several scientists, each specializing in one of the pathogens selected for the BioHealthBase BRC, serve as co-investigators for the project.

Name	Affiliation	Organism
Dr. Hilary Morrison	Marine Biological Laboratory	<i>Giardia lamblia</i>
Dr. Stephen Johnston	Arizona State University	<i>Mycobacterium tuberculosis</i>
Dr. Adolfo Garcia-Sastre	Mount Sinai School of Medicine	<i>Influenza virus</i>
Dr. Barbara J. Mann	University of Virginia	<i>Francisella tularensis</i>
Dr. Louis M. Weiss	Albert Einstein College of Medicine of Yeshiva University	<i>Microsporidia</i>
Dr. Ellen S. Vitetta	University of Texas Southwestern Medical Center	<i>Ricinus communis</i>

Scientific Working Group (SWG)

The Scientific Working Group (SWG) will serve as a steering committee of scientific experts from outside the project development group who will advise the BioHealthBase BRC team regarding system requirements and system utility. The SWG has been selected to represent the user community based on their research interests, research productivity and interest in database development and data analysis. The goal will be to use the SWG as a sounding board to validate perceived system requirements and to assist in the testing of system components. The SWG will also serve as a source of information concerning the members of the relevant scientific community.

Name	Affiliation	Research Interests
Catherine Macken	Los Alamos National Labs	<i>Influenza virus</i>
Robin Bush	University of California, Irvine	<i>Influenza virus</i>
Patrick Keeling	University of British Columbia	<i>Microsporidia</i>
John Samuelson	Boston University	Protozoan parasites
John McKinney	The Rockefeller University	<i>Mycobacterium tuberculosis</i>
Megan Murray	Harvard University	<i>Mycobacterium tuberculosis</i>
Kevin McIver	University of Texas Southwestern Medical Center	<i>Francisella tularensis</i>
Harold "Skip" Garner	University of Texas Southwestern Medical Center	Computational biology and bioinformatics
Warren Gish	Washington University	Computational biology and bioinformatics

Internet